EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XIX, NO. 1, 1959

SIGNIFICANCE LEVELS FOR THE KUDER-RICHARDSON (21) RELIABILITY COEFFICIENT

SAMUEL B. LYERLY
Society for the Investigation of Human Ecology

IN the development of a new test, particularly one based upon a new principle or designed to test a special experimental hypothesis, it may be worth while to try out a small number of items on a small sample of subjects to find out whether the test idea has sufficient merit to encourage further development. Data gathered in such a "pilot" study may then be analyzed by studying such of their characteristics as are pertinent to the investigation. An estimate of reliability may be made, and it may turn out to be low—so low that one wonders whether it is significantly different from zero. A test of the null hypothesis will help the investigator decide whether to abandon the project or to continue development of the test by revising, adding items, etc.

If a retest or if parallel forms have been properly administered, a product-moment correlation coefficient can be computed, and the t test or published tables of the significant values of r will provide the answer. If a Kuder-Richardson reliability estimate is made, however, the standard significance test is not appropriate.

Case IV in the Kuder-Richardson formulation, Equation 21 of their 1937 paper (4), may be written

$$[1] \qquad \text{K-R 21} = \frac{n}{n-1}\left[1 - \frac{\bar{X} - \bar{X}^2/n}{s^2}\right],$$

where n is the number of items in the test, $\overline{X}$ is the mean of the sample of scores, and s^2 is the variance of the scores. If the fraction on the right in the parentheses is inverted and then multiplied by N, where N is the number of cases in the sample, we have a quantity whose distribution approaches that of χ^2 for $N-1$ degrees of freedom. Most of the standard statistics texts, such as (1) pp. 138–139, (2) pp. 231–232, (3)

pp. 195–197, describe this statistic. It is frequently called the "binomial index of dispersion", and is used to test the hypothesis that N observed frequencies of a binomial variable, each based upon n trials, are homogeneous. As applied to mental tests, the corresponding null hypothesis is that the observed numbers of successes at n test items scored 1 or 0 does not vary among the sample of N subjects more than would be expected if their abilities were all equal.

Writing Kuder-Richardson formula 21 and χ^2 in terms of each other, we have

$$[2] \qquad \chi^2 = \frac{nN}{n - (n-1)\,\text{K-R 21}}$$

$$[3] \qquad \text{K-R 21} = \frac{n}{n-1}\left[1 - \frac{N}{\chi^2}\right] .$$

From these equations several interesting relationships can be inferred, but we pass on to the significance problem. Using [3], the one-tail .05, .01, and .001 points of K-R 21 were computed for some representative values of N and n. Published tables of χ^2 were used for $N \leqq 30$, and for a higher N, Fisher's normal approximation was used, $\sqrt{2\chi^2} - \sqrt{2N-3}$. (The quantity under the second radical is one less than twice the number of degrees of freedom.) Table 1 lists these significance levels. The table is presented for illustrative purposes rather than for reference; since the statistic is so easy to compute, an investigator can readily determine the value appropriate for his own particular test length and sample size.

From the table we see that the number of subjects affects the size of the significant values more than does the number of items. It is interesting (and encouraging) to note also that for small values of N numerically *smaller* values of K-R 21 are required for significance than is the case with a product-moment coefficient computed for a sample of the same size. This is a consequence of the fact that the null distributions of these statistics are quite different. The distribution of r is symmetric about zero, while that of K-R 21 is skewed and zero is not the most probable value (mode), the median (50 per cent point), nor the expectation (mean); these three values are all different from each other, and all negative. There is an upper limiting value of 1.0, but no limiting negative value above $-\infty$. Thus K-R 21 and r have different scales and are not strictly comparable.

TABLE 1

*Values of Kuder-Richardson Formula 21 Significant at the .05, .01, and .001 Levels**

		n					
N	P	10	20	30	40	50	100
	.05	.455	.431	.423	.420	.418	.413
10	.01	.599	.567	.557	.553	.550	.544
	.001	.713	.675	.664	.658	.655	.648
	.05	.408	.386	.380	.377	.374	.371
15	.01	.540	.511	.502	.498	.496	.491
	.001	.650	.616	.605	.600	.597	.591
	.05	.374	.355	.349	.346	.344	.340
20	.01	.498	.471	.463	.459	.457	.452
	.001	.604	.573	.563	.558	.555	.550
	.05	.349	.330	.325	.322	.320	.317
25	.01	.465	.441	.433	.429	.427	.423
	.001	.569	.539	.530	.525	.522	.517
	.05	.328	.311	.306	.303	.302	.299
30	.01	.439	.416	.409	.406	.404	.399
	.001	.540	.511	.503	.498	.496	.491
	.05	.293	.277	.273	.270	.269	.266
40	.01	.390	.370	.363	.360	.358	.355
	.001	.480	.455	.447	.443	.441	.437
	.05	.270	.256	.252	.250	.248	.246
50	.01	.362	.343	.337	.334	.332	.329
	.001	.448	.424	.417	.413	.411	.407
	.05	.208	.197	.193	.192	.191	.189
100	.01	.281	.267	.262	.260	.259	.256
	.001	.354	.335	.330	.327	.325	.322

* The third digit has been *raised* to the next higher value so that the entries are all at or above the indicated level.

REFERENCES

1. Anderson, R. L. and Bancroft, T. A. *Statistical Theory in Research.* New York: McGraw-Hill Book Company, 1952.
2. Dixon, W. J. and Massey, F. J. *Introduction to Statistical Analysis.* (Second Edition). New York: McGraw-Hill Book Company, 1957.
3. Hoel, P. G. *Introduction to Mathematical Statistics.* New York: John Wiley and Sons, 1947.
4. Kuder, G. F. and Richardson, M. W. "The Theory of the Estimation of Test Reliability." *Psychometrika,* II (1937), 151–160.